International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

Business Analytics Using Predictive Algorithms

Dhanshri Satish Jangam[1], Dr. Arati R. Deshpande[2]
[1]Department of Computer Engineering
SCRT's Pune Institute of Computer Technology
Pune, India
dhanshrijangam5@gmail.com
[2]Department of Computer Engineering
SCRT's Pune Institute of Computer Technology
Pune, India
ardeshpande@pict.edu

Abstract— In today's data-driven business landscape, organizations strive to extract actionable insights and make informed decisions using their vast data. Business analytics, combining data analysis, statistical modeling, and predictive algorithms, is crucial for transforming raw data into meaningful information. However, there are gaps in the field, such as limited industry focus, algorithm comparison, and data quality challenges. This work aims to address these gaps by demonstrating how predictive algorithms can be applied across business domains for pattern identification, trend forecasting, and accurate predictions. The report focuses on sales forecasting and topic modeling, comparing the performance of various algorithms including Linear Regression, Random Forest Regression, XGBoost, LSTMs, and ARIMA. It emphasizes the importance of data preprocessing, feature selection, and model evaluation for reliable sales forecasts, while utilizing S-BERT, UMAP, and HDBScan unsupervised algorithms for extracting valuable insights from unstructured textual data.

Keywords- Business Analytics; Dashboards; Sales Forecasting; Topic Modelling; Machine Learning; Natural Language Processing.

I. INTRODUCTION

Business analytics is the practice of using data analysis, statistical modeling, and other quantitative techniques to extract meaningful insights from data and support decision-making in business contexts. It involves collecting, processing, analyzing, and interpreting data to identify trends, patterns, and relationships that can drive strategic planning, operational optimization, and overall business performance.

Big Data has been widely adopted in place of conventional methods to manage various business operations and create more accurate predictive models for the organization. Many businesses are enhancing their effectiveness in delivering client pleasure thanks to business intelligence and analytics. One of the key factors in the recent radical changes in the goods and services that businesses offer is predictive modelling [21].

Business intelligence (BI) encompasses a wide range of business-centric practices and methodologies that can be used in e-governance, healthcare, e-commerce, security, and market intelligence, among other applications [19]. Different tools, offerings, and technologies are needed for BI, which is frequently used in supply chain, sales, finance, and marketing [18]. The evolution of business analytics can be attributed to the advancements in technology, increased data availability, and the growing complexity of business operations. Initially, businesses relied on descriptive analytics, which focused on summarizing historical data to gain a better understanding of past performance. Business analytics using traditional methods involves the application of techniques such as descriptive analytics, inferential statistics, trend analysis, correlation analysis, comparative analysis, and decision trees to gain insights and support decision-making in business contexts. Descriptive analytics provided valuable insights into what happened in the past and helped organizations track key performance indicators. Various analyses are carried out by data analysts using one or more predictive analytics modelling tools. To create better predictive models, data analysts use a variety of machine learning algorithms as well as other statistical methods [20]. By using a variety of different algorithms to the data at hand, predictive analytics seeks to identify patterns that can help forecast future behavior. Predictive analytics is an ongoing process because if the model outputs are discovered to accurately forecast anything, the business will try to come up with a different solution to prevent customers from leaving their network [22].

Sales forecasting and topic modeling are two important applications within the domain of business analytics. By leveraging predictive algorithms, businesses can identify patterns, make accurate predictions, and uncover hidden insights from their data. This enables organizations to optimize their operations, enhance customer experiences, mitigate risks, and drive business growth. Predictive analytics algorithms such

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

as regression analysis, decision trees, random forests, XGBoost, LSTMs (Long Short-Term Memory), and ARIMA (Auto Regressive Integrated Moving Average) time series forecasting are commonly used in business analytics to achieve accurate predictions and valuable insights. Business analytics is a powerful discipline that harnesses data to gain insights, make informed decisions, and drive business success.

The evolution of business analytics has seen a shift from descriptive analytics to predictive analytics, enabling businesses to anticipate future trends and behaviors. Sales forecasting and topic modeling are two significant applications within business analytics, while business analytics using predictive algorithms empowers organizations to make accurate predictions and optimize various aspects of their operations. By leveraging the potential of business analytics and predictive algorithms, organizations can stay competitive, adapt to market dynamics, and unlock new opportunities for growth and innovation.

II. RELATED WORK

Ref. No.	Literature Survey Analytical Table			
	Description	***Methodology Used***	***Dataset***	***Pros and Cons***
[1]	In this study, machine learning algorithms are used to a dataset in the Hadoop MapReduce environment to identify missing values and identify trends [1].	Decision Tree C4.5, Missing Values Imputation (MVI) algorithm	Pima Indian diabetes	The proposed methods can detect trends in the dataset and impute missing values [1].
[2]	A technique that employs predictive machine learning algorithms to forecast users' gender and age based on their actions, services, and contract information [2].	k-Means clustering, PCA, SVM, XGBoost, Random forest	Cells and sites database, Call Detail Records	Based on information from mobile phones, there has been an improvement in the precision of user age and gender predictions [2].
[3]	The goal of this study is to create a prediction system using machine learning and choose the best classifier to produce results that are most similar to clinical outcomes. The suggested approach tries to concentrate on choosing the characteristics that fail in applying predictive analysis to identify Diabetes Miletus early on [3].	Random forest, Decision Tree, SVM, Naive Bayes, K nearest neighbour (KNN)	Diabetes Dataset	The categorization technique is now more accurate, which will speed up the forecasting process. Techniques for assembling can be employed to improve performance [3].
[4]	This study uses a machine learning supervised algorithm for classification to analyse the potential cross-selling of telecom customers [4].	Naive Bayes Classier, C4.5	PT. Telkom Jakarta	Cross-selling marketing strategies with data mining methods employing the C4.5 algorithm, which has an excellent accuracy value of 88.61% and AUC 0.870 [4].
[5]	Using five models for assessing bankruptcy risk that are well-known in the literature - Altman, Conan and Holder, Tafler, Springate, and Zmijewski - the goal of this work was to increase knowledge of how to predict a company's bankruptcy and analyse the predictive power of factor analysis based over discriminant analysis [5].	Principal Component Analysis	Amadeus database	Higher performance is the key to lowering bankruptcy risk, according to the PCA scoring methodology used to derive bankruptcy risk ratings by year and nation using discriminant analysis of indicators of bankruptcy risk [5].
[6]	To stop unauthorised access to healthcare data, this analysis method takes into account HSs powered by smart grids. Based on the	PDA-HS (Predictive Data Analysis)	10,000 synthetic data records of the Patients	For various users, accuracy increased by 4.1%, while processing time and data loss decreased by 3.42% and 13.95%,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

Ref. No.	Literature Survey Analytical Table			
	Description	***Methodology Used***	***Dataset***	***Pros and Cons***
	information provided with the user in various contexts, access was examined [6].			respectively [6].
[7]	A web-based tool for stock prediction has been created using machine learning algorithms, and it consists of five modules: a dashboard, individual stock prediction pages, sentiment analysis results, trade assistant chatbot, and trading site [7].	Long Short Term Memory, SVM, Random Forest	NSEtools, NSEpy, Quandl	According to the results of repeated hyperparameter adjustment, the model was operating with an accuracy ranging from 95% to 98% [7].
[8]	This study offers a soft computing-based stacking ensemble classification model for detecting credit card fraud [8].	Fuzzy Nearest Neighbour(FNN), Fuzzy-Rough Nearest Neighbour(FRNN)	Credit Card Fraud from UCI	Using a 10-fold cross validation technique, a detection rate of 84.90% and an AUC of 0.8555 are produced for the Australian credit approval dataset, and a detection rate of 76.30% and an AUC of 0.6795 are produced for the German credit dataset [8].
[9]	The suggested method is made up of three primary steps: processing data and discretization, looking for related patterns using FP-Growth, and identifying and analysing fraud patterns [9].	Association Rule Mining	Financial statement	The usage of companies registered on the Thai Stock Exchange without taking into account the nature of their businesses was one of the limitations of this study [9].
[10]	The study shows a useful system that makes use of	Rule Induction (RI), Naive	Statlog (German	The real payment card database usage is

Ref. No.	Literature Survey Analytical Table			
	Description	***Methodology Used***	***Dataset***	***Pros and Cons***
	aggregated features for payment card fraud detection and real transaction records for evaluation and demonstration of the proposed system's performance [10].	Bayes, Decision Trees(DT), Decision Stump (DS), SVM	Credit), Statlog (Australian Credit), and Default Credit Card.	limited to a financial institution in Malaysia [10].
[11]	This proposed work presents a predictive maintenance system for manufacturing environments based on machine learning [11].	XGBoost, Gradient Boosting and AdaBoost, MLP Regressor, SVR, Multiple Regression, Lasso Regression and Ridge Regression.	Collected from IoT sensors	Digitization has taken over the production process [11].
[12]	By applying machine learning-based framework and techniques, this study examines the site choice made by Chinese businesses in the global network. These include the hierarchical cluster analysis, heat maps, and the 3D vision of mode networks using the internet's intensity as a stand-in for the Fourth Industrial Revolution [12].	Neural Network	Wind, World Bank,	Increased financial leverage has a cost that may outweigh any possible advantages of global expansion [12].
[13]	In order to extract the energy usage based on previous data, predictive analysis had been performed on the	Deep Learning Regression	Energy consumption	The outcome of the predictive analysis aids in smart manufacturing and the preparation of the company's

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

Ref. No.	Literature Survey Analytical Table			
	Description	***Methodology Used***	***Dataset***	***Pros and Cons***
	values of a smart manufacturing organisation [13].			advanced budget in numerous ways [13].
[14]	A new method that takes into account the distinct behaviour of process activities for predicting completion times and analysing performance in manufacturing. To enable the creation of a probabilistic model and predictive models, the framework discusses process mining techniques [14].	Bayesian Networks	Event log	The system demonstrates that users can validate it in a simple and reliable manner, ensuring that any mistakes made during construction are corrected before it is used in practical applications [14].

III. METHODOLOGY

Figure 1 describes the proposed system architecture below.

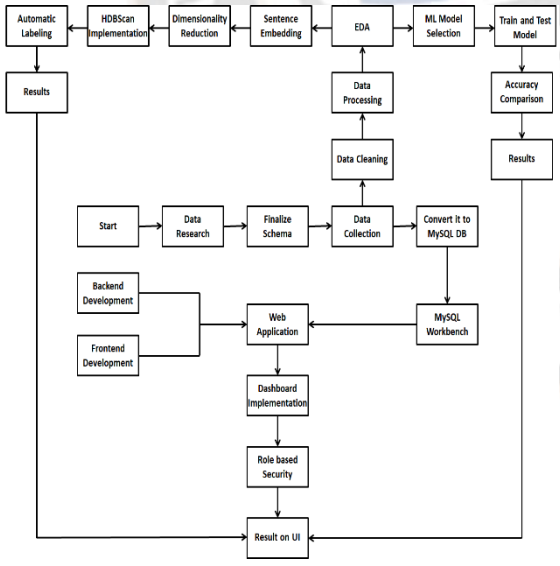

Figure 1. System Architecture.

A. Data Research

Fact Constellation schema has been finalized after successful Industry research. The finalized schema comprises 17 tables that include OIC (Interaction Company), OP (Product), OCR (Cancellation Revenue), OCCS (Company Current Summary), OICT (Interaction Contact), OTS (Tech-support), OT (Touch point), OIPL (Invoice Product List), OI (Invoice), OCN (Company Name), OC (Contact), OCA (Contact Address), OCE (Contact Email), OCIP (Contact Industry Pit), OCP (Contact Phone), OCU (Contact URL), and OS (Surveys). These tables form the structure for organizing and storing data related to various aspects of the company's interactions, products, revenue, customer contacts, invoices, and surveys.

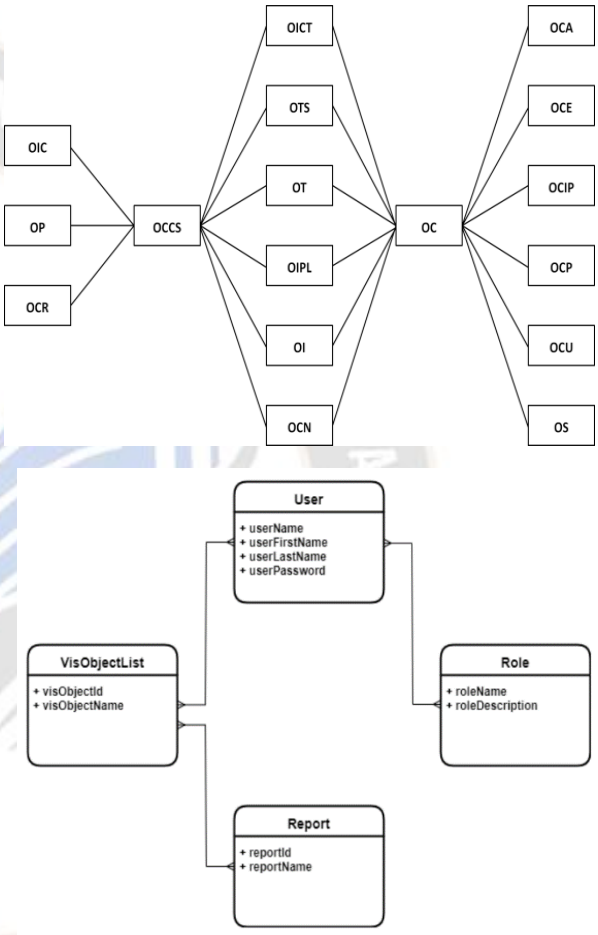

Figure 3. Fact Constellation schema for Dataset.

B. Data Collection

The data for the schema has been gathered through a combination of sources. Firstly, the Faker library, an open-source tool for generating realistic and diverse datasets, was utilized to generate synthetic data that aligns with the schema's structure. Secondly, an open-source dataset was employed to supplement the information obtained from Faker, providing additional real-world data points. Lastly, to enhance the accuracy and completeness of the collected data, manual data collection methods were employed, involving direct input from individuals or extraction from reliable sources. The combination of these approaches ensures a comprehensive and diverse dataset that aligns with the finalized schema.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

C. Module I

1) Data Processing

Following the data collection stage, the accumulated data from various sources, including the Faker library, open-source dataset, and manual collection efforts, exists in the form of CSV files. To further process and organize this data, it has been transformed into a MySQL database format using the SQL Alchemy library. SQL Alchemy is a powerful Python library that provides an abstraction layer for working with databases. By utilizing SQL Alchemy, the CSV files has been imported and converted into tables within a MySQL database, facilitating efficient data storage, retrieval, and manipulation.

2) Design and Planning

During the design and planning phase, careful consideration has been given to the various sub modules that constitute the web application. These sub modules comprise the login page, register new user, and create/edit/view/delete reports functionalities. The login page sub module has been designed with a focus on security and usability, providing a seamless and intuitive interface for users to authenticate and access the application. Lastly, the create/edit/view/delete reports sub module allows users to generate, modify, view, and delete reports, providing a comprehensive and efficient reporting system within the web application.

3) Backend Development using Spring boot

This phase involved the creation of four entities: User, Role, VisObjectList, and Report. The application development process included the implementation of APIs for various operations, such as GET, POST, DELETE, and PATCH, which interact with a MySQL database.

The User entity represents users of the application, while the Role entity defines different roles and permissions for users. The VisObjectList entity manages visual object lists, and the Report entity handles the generation and management of reports. The APIs developed enable communication between the frontend and backend, allowing users to retrieve, create, delete, and update data stored in the MySQL database.

Figure 4. ER Diagram for Backend.

4) Frontend Development using Angular

During the frontend development phase using Angular, various user interface (UI) components have been implemented. These components include the UI for Register New User, Login, Create Customized Report, View Report Details, View/Edit/Delete Report, View/Edit VisObjects Details per user, and View Reports. The UI for Register New User allows individuals to create new accounts, providing a user-friendly registration process. The Login UI enables users to authenticate and access the application securely.

The Create Customized Report UI allows users to generate personalized reports based on their requirements. The View Report Details UI provides a comprehensive view of the report's contents. The View/Edit/Delete Report UI permits users to interact with reports, offering options to modify, delete, or edit report details. The View/Edit VisObjects Details per user UI allows users to manage and customize visual objects associated with their reports.

Lastly, the View Reports UI provides a convenient interface for users to browse and access existing reports. These UI components collectively enhance the user experience and facilitate efficient utilization of the web application.

5) Dashboard Implementation

As part of the dashboard implementation phase, the VisObjects have been created utilizing the Chart.js library. Chart.js is a JavaScript library renowned for its capabilities in developing interactive and visually appealing charts and visualizations. The VisObjects have been purposefully designed to present data in a meaningful and engaging manner. Leveraging the functionalities offered by Chart.js, an array of chart types including line charts, bar charts, pie charts, and more have been integrated into the dashboard.

By incorporating Chart.js in the dashboard, users are provided with comprehensive visual representations of their data. This empowers them to easily interpret and analyze information, facilitating informed decision-making and delivering an enhanced user experience.

6) Role Based Security

The Role Based Security implementation in this system includes two roles: user and admin. The admin role possesses the authority to Create Customized Reports, View Report Details, View/Edit/Delete Reports, and View/Edit VisObjects Details per user. On the other hand, users have the privilege to View Report Details and View Reports, but only for the ones they have been granted access to by the admin.

The admin is responsible for assigning permissions to users. If a user lacks permission to view specific VisObjects, they will be restricted from accessing them. This ensures that users can only view the VisObjects for which they have been granted authorization, maintaining data privacy and security within the application.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

D. Module II

1) Preprocessing

During the data preprocessing phase of Module II, several cleaning techniques have been applied to ensure the quality and consistency of the data.

Figure 5. Word Cloud for Module II Dataset.

These cleaning steps include removing emojis, eliminating hyperlinks and markup, discarding numeric values, unifying whitespaces, removing punctuation marks, eliminating stop words, and stemming words.

The removal of emojis helps in simplifying the text and removing any non-essential graphical elements. Hyperlinks and markup are eliminated to focus solely on the textual content. Numeric values are removed to maintain consistency in the data.

Unifying whitespaces involves reducing multiple consecutive whitespaces to a single space, enhancing the readability of the text. Punctuation marks are removed to extract only the essential words.

Additionally, the elimination of stop words, such as common words like "the" or "is," aids in removing noise from the text. Finally, stemming words reduces words to their base form, ensuring consistency and simplifying analysis.

	app_id	review_text	review_score	review_votes	review_text_clean
238958	365800	I don't really like this game, It's very repetitive and not very user friendly.	0	0	realli like game repetit user friend
267206	409670	Hated the controls	0	0	hate control
202994	313630	Very pretty, but needs a lot more work as a VR title. The moment system is very clunky with not enough moter control options. It feels as if you're fighting with the game rather than playing it. Hopeful this will be patched. If you're buying this because it's another game with VR support, skip it until it gets more patching. ## VR System Used: Vive ##	0	0	pretti need lot work vr titl moment system clunki enough moter control option feel fight game rather play hope patch buy anoth game vr support skip get patch # # vr system use vive # #
318271	92100	4/10 - better take a detour around this game Neat idea, neat presentation, neat graphics.....and the same as with all things 'neat' - they get boring pretty quickly.	0	1	/ - better take detour around game neat idea neat present neat graphicsand thing neat ' - get bore pretti quick
24812	202130	Just ok... lots of repetition and really I recommend you go play Dungeon Keeper for the real experiance. Even with 1997 graphics the origional Dungeon Keeper beats this for game play and enjoyment!!	0	0	ok lot repetit realli recommend go play dungeon keeper real experi even graphic origion dungeon keeper beat game play enjoy

Figure 6. Dataframe for Cleaned Preprocessed Data.

2) Processing

In the data processing phase, the focus is on negative reviews. Only rows where the review score equals zero are considered, and further filtering is done by selecting sentences with a length of more than four tokens. This ensures that only the relevant and meaningful negative reviews are included for further analysis and processing.

	app_id	review_text	review_score	review_votes	review_text_clean
238958	365800	I don't really like this game, It's very repetitive and not very user friendly.	0	0	realli like game repetit user friend
267206	409670	Hated the controls	0	0	hate control
202994	313630	Very pretty, but needs a lot more work as a VR title. The moment system is very clunky with not enough moter control options. It feels as if you're fighting with the game rather than playing it. Hopeful this will be patched. If you're buying this because it's another game with VR support, skip it until it gets more patching. ## VR System Used: Vive ##	0	0	pretti need lot work vr titl moment system clunki enough moter control option feel fight game rather play hope patch buy anoth game vr support skip get patch # # vr system use vive # #
318271	92100	4/10 - better take a detour around this game Neat idea, neat presentation, neat graphics.....and the same as with all things 'neat' - they get boring pretty quickly.	0	1	/ - better take detour around game neat idea neat present neat graphicsand thing neat ' - get bore pretti quick
24812	202130	Just ok... lots of repetition and really I recommend you go play Dungeon Keeper for the real experiance. Even with 1997 graphics the origional Dungeon Keeper beats this for game play and enjoyment!!	0	0	ok lot repetit realli recommend go play dungeon keeper real experi even graphic origion dungeon keeper beat game play enjoy

Figure 7. Dataframe for Preprocessed Data.

3) Sentence Embedding

Sentence embedding refers to the process of representing a sentence or a piece of text as a fixed-length numerical vector. It aims to capture the semantic meaning and contextual information of the sentence in a compact and meaningful way. Sentence embeddings are derived using various techniques such as neural networks, word embeddings, or pre-trained models. These embeddings enable comparison,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

similarity measurement, and analysis of sentences in a numerical space, facilitating various natural language processing tasks like text classification, sentiment analysis, and information retrieval. By converting sentences into dense vector representations, sentence embedding allows machines to understand and reason about textual information more effectively.

	0	1	2	3	4	5	6	7	8
0	0.040145	-0.004369	-0.037483	0.062973	-0.018969	-0.028973	0.008283	-0.020792	-0.029848
1	0.020373	0.023936	-0.023920	0.018118	-0.022435	0.041810	-0.019868	-0.015140	0.043081
2	-0.009930	0.029309	-0.010856	-0.005844	0.037529	-0.013164	-0.037438	0.063740	-0.031900
3	0.012923	-0.027822	-0.015407	0.078636	-0.006286	-0.003896	-0.060726	0.012985	-0.052563
4	0.024937	0.000918	-0.042397	0.082932	-0.004724	0.011544	-0.014959	0.023656	-0.016062

5 rows × 768 columns

Figure 8. Sentence Embeddings.

4) Dimesionality Reduction

Dimensionality reduction is a technique used to reduce the number of features or variables in a dataset while preserving the essential information. It aims to simplify complex data by transforming it into a lower-dimensional space, removing redundant or less important features. This helps in visualizing and analyzing data, improving computational efficiency, and mitigating the curse of dimensionality.

5) Clustering

Clustering is an unsupervised machine learning technique that groups similar data points together based on their inherent characteristics. It aims to discover underlying patterns or structures within the data without prior knowledge of class labels. The algorithm assigns data points to clusters such that points within the same cluster are more like each other compared to points in different clusters. Clustering finds applications in various domains, such as customer segmentation, image segmentation, anomaly detection, and recommendation systems. It helps in organizing and understanding large datasets, identifying cohesive groups, and uncovering hidden insights in the data.

E. Module III

1) Preprocessing

First, the load_data() function reads the sales data from a CSV file specified by the 'train.csv' URL and returns it as a pandas Data frame.

Second, the monthly_sales(data) function takes the sales data as input, processes it, and performs the following steps:

a) Step 1: Removes the day indicator from the date column to retain only the year and month information.

b) Step 2: Calculates the total sales for each month by grouping the data based on the modified date column and summing the sales values.

c) Step 3: Converts the modified date column to a datetime format.

d) Step 4: Saves the resulting monthly sales data as a new CSV file named 'monthly_data.csv'.

```
sales_data.head()
```

	date	store	item	sales
0	2013-01-01	1	1	13
1	2013-01-02	1	1	11
2	2013-01-03	1	1	14
3	2013-01-04	1	1	13
4	2013-01-05	1	1	10

```
monthly_data.head()
```

	date	sales
0	2013-01-01	454904
1	2013-02-01	459417
2	2013-03-01	617382
3	2013-04-01	682274
4	2013-05-01	763242

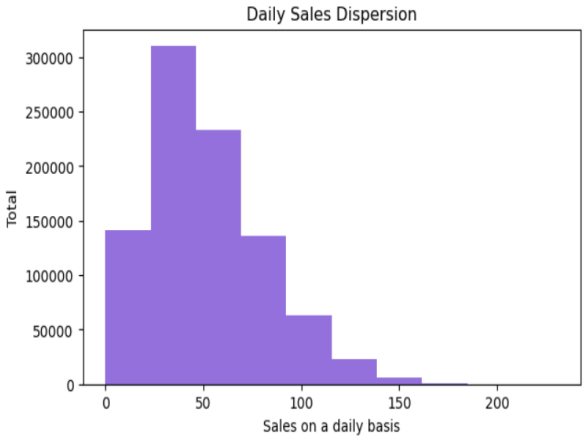

Figure 9. Dataset for Module III.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

2) Exploratory Data Analysis (EDA)

a) Daily Sales Dispersion

The term "daily sales dispersion" describes the variance or spread of sales numbers from day to day within a specific time frame. It gauges how much sales vary or differ from the normal or anticipated level of sales. Businesses may evaluate and manage risks, optimize inventory, and create powerful sales strategies with the aid of an understanding of daily sales dispersion.

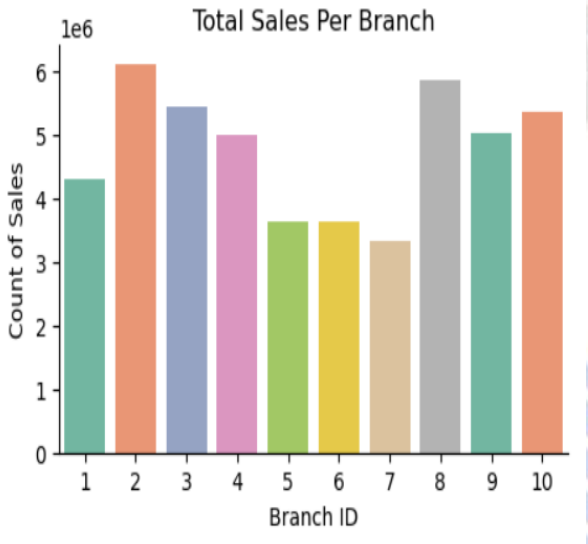

Figure 10. Daily Sales Dispersion.

b) Total Sales Per Branch

The total sales produced by a specific branch or location of a firm over the course of a specific time period are referred to as total sales per branch. It indicates the total income or revenue made by that branch. Businesses can assess the efficiency and profitability of individual branches, pinpoint locations that perform well or poorly, allocate resources wisely, and make well-informed decisions about branch network expansion, consolidation, and restructuring by analyzing total sales per branch.

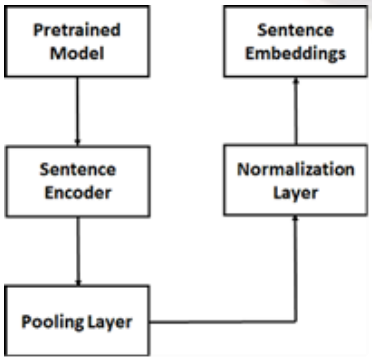

Figure 11. Total Sales Per Branch.

3) Modeling

Modeling in supervised machine learning involves creating a predictive model that learns from labeled training data to make accurate predictions or classifications on unseen data. It follows a supervised learning paradigm where the model is provided with input features and corresponding target labels during training. The modeling process includes selecting an appropriate algorithm (e.g., LSTM, Linear Regression, XGBoost, and ARIMA) that can capture the relationships between input features and target labels. The model is trained by adjusting its internal parameters to minimize the prediction errors. Once trained, the model can be used to make predictions on new data by providing the input features. The performance of the model is evaluated using metrics like RMSE, MAE and F2 score to assess its effectiveness in making accurate predictions.

IV. MODEL

A. SBERT (Sentence Bidirectional Encoder Representations from Transformers)

Being a "twin network," SBERT may analyze two sentences concurrently and in a similar manner. Human ability to think of this design as a single model applied many times is made possible by the fact that these two twins are similar in every way (their weights are coupled). Sentence-BERT is an architecture that builds embeddings (vector representations) for sentences using pre-training and fine-tuning techniques. Such embeddings comprise the phrases' conceptual meaning and can be applied to several natural language processing (NLP) tasks, including classification, clustering, and sentence similarity.

An overview of each element of the SBERT architecture is given below:

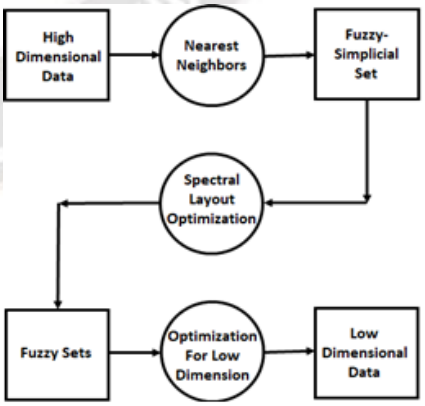

Figure 12. Block Diagram of SBERT.

1) Pre-trained Models

These are models that have already been trained using methods like unsupervised learning on significant volumes of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

textual data. The pre-trained models are applied by the sentence encoder component.

2) Sentence Encoder

The sentence encoder transforms input sentences into high-dimensional representations. Typically, it bases itself on pre-trained models. Sentence tokens are processed by the encoder, which produces contextualized embeddings as a result. The function E_enc can be used to represent an encoder.

$$EF: E\_enc(S_i) \rightarrow CTE. \quad (1)$$

Where,

EF – Encoder Function.

E_enc – An Encoder.

S_i – Input in the form of Sentences.

CTE – Contextualized Embeddings.

3) Pooling Layer

It collects the series of contextualized embeddings generated by the sentence encoder to build a fixed-size representation of the complete sentence. Mean pooling and maximum pooling are two frequent pooling techniques.

$$PF: E\_pool(CTE) \rightarrow PRP. \quad (2)$$

Where,

PF – Pooling Function.

E_pool – Pooling Layer.

CTE – Contextualized Embeddings.

PRP – Pooled Representation.

4) Normalization Layer

To guarantee that the sentence embeddings have a constant scale and are simple to compare across various sentences, the output from the pooling layer is routed via a normalization layer.

$$NF: E\_norm(PRP) \rightarrow NME. \quad (3)$$

Where,

NF – Normalization Function.

E_norm – Normalization Layer.

PRP – Pooled Representation.

NME – Normalized Embeddings.

5) Sentence Embeddings

After the normalization layer, these are the final representations of the sentences. These embeddings provide a dense vector representation of the sentences' semantic meaning.

$$STE: E_i = NME. \quad (4)$$

Where,

STE – Sentence Embeddings.

E_i – A dense vector representation of the sentences' semantic meaning.

NME – Normalized Embeddings.

B. UMAP (Uniform Manifold Approximation and Projection)

An overview of each element of the UMAP block diagram is given below:

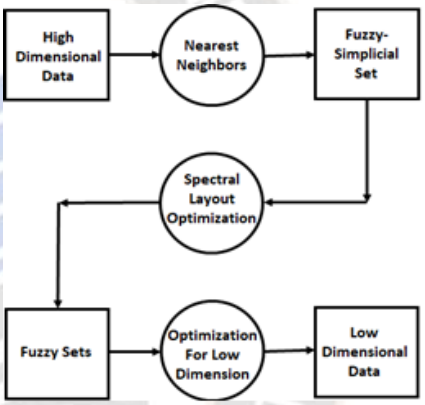

Figure 13. Block Diagram of UMAP.

1) High-dimensional Data

This reflects the initially collected data in a high-dimensional space, generally shown as a feature matrix.

2) Nearest Neighbors

Using the data that is high-dimensional, UMAP first creates a k-nearest neighbor graph. Each point in the dataset has its closest neighbors identified. The value k, which represents the number of closest neighbors, can be adjusted.

a) Let X be the high-dimensional data matrix.

b) Compute pairwise distances between data points:

$$D = distance\_matrix(X, X). \quad (5)$$

c) For each data point x_i, find the k nearest neighbors based on D.

3) Fuzzy-Simplicial Set

The nearest neighbor graph is transformed into a fuzzy topological structure known as a fuzzy-simplicial set

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

using UMAP. In this stage, each point is given a level of membership based on how connected it is to its neighbors. To gauge the degree of the link, it uses a membership function.

Construct a fuzzy membership matrix F, where F[i, j] represents the membership strength between data points x_i and x_j. Fuzzify the nearest neighbors by applying a membership function, such as the fuzzy set membership function.

4) Spectral Layout Optimization

A low-dimensional network that resembles the high-dimensional data can be generated using the fuzzy-simplicial set. UMAP optimizes the configuration of this graph in the low-dimensional space using spectral approaches.

a) Construct an adjacency matrix A based on F.

b) Compute the graph Laplacian matrix

$$L = D - A. \quad (6)$$

Where,

D is the diagonal matrix of row sums of A.

c) Perform eigen decomposition of L to obtain the eigenvectors and eigenvalues.

d) Select a subset of eigenvectors corresponding to the smallest eigenvalues.

e) Embed the data points in the low-dimensional space using the selected eigenvectors.

5) Fuzzy Sets

To select the best way to represent the low-dimensional graph, UMAP makes use of fuzzy set theory. Each data point in the low-dimensional space is given a fuzzy membership value showing whether it belongs to a particular cluster or group.

6) Optimization for Low Dimension

To identify an embedding that reduces the difference between the high-dimensional and low-dimensional graphs, UMAP performs optimization in the low-dimensional space. By balancing preserving the integrity of close points and the general pattern of the data, it seeks to maintain both local and global structure.

a) Minimize the discrepancy between the high-dimensional and low-dimensional affinities.

b) Use stochastic gradient descent or other optimization techniques to find an embedding that preserves both local and global structure.

7) Low-dimensional Data

The original high-dimensional data's low-dimensional representation is the UMAP algorithm's ultimate output. This generally two- or three-dimensional representation can be visualized to show the data's patterns and organizational structure.

C. HDBScan

An overview of each element of the HDBScan block diagram is given below:

1) Input Data

This is the initial dataset comprising the clustering candidate points.

2) Core Distance Computation

Calculate the core distance for each point in the dataset using the core distance computation method. The distance between a point's core and its kth nearest neighbor serves as a gauge of that point's local density.

3) Mutual Reachability Distance

Determine each pair of points' mutual reachability distance by using the formula below. The reciprocal reachability distance is a metric of resemblance or connectedness between places that considers both their proximity and their core distances.

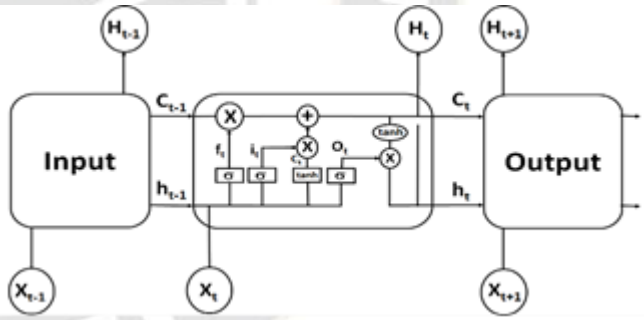

Figure 14. Block Diagram of HDBScan.

4) Minimum Spanning Tree Construction

Using the mutual reachability distances, construct the minimum spanning tree (MST). The MST links locations according on how easily they may be reached from one another, creating a hierarchical structure.

5) Cluster Tree Construction and Condensation

Constructing a cluster tree from the MST and condensing it entails recursively dividing clusters depending on the connectivity of points. At various levels of the structure, clusters are merged and condensed in this process.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

6) Clustering Outcome

Based on user-defined criteria, such as a threshold value or a minimum cluster size, extract the clusters from the cluster tree. Give the points in the dataset cluster labels.

7) Outliers and Noise Points

Identify outliers and noise points that do not fit into any cluster. These are the locations that do not satisfy the connection or minimum cluster size requirements.

8) Final Results – Cluster Labels

Assign noise and outlier class labels to the clusters of noise points and outliers.

9) Clustered Data

The dataset with assigned cluster labels is the algorithm's final output, and each point is either allocated a label for the cluster it belongs to or is designated as noise or an outlier.

D. Long Short Term Memory (LSTM)

The LSTM's first step is to decide which data from the cell state will be thrown out. This determination is made by the sigmoid "forget gate layer." Each number in the cell state C_{t-1} receives an evaluation of the values in h_{t-1} and a return value between 0 and 1. The cell state, represented by the horizontal axis running through the top of the picture, is crucial to long-term short-term memory.

An overview of each element of the LSTM block diagram is given below:

Figure 15. Block Diagram of LSTM.

1) Forget Gate

The forget gate purges the data that is no longer relevant in the cell state. The gate receives two inputs, x_t (input at the current time) and h_{t-1} (prior cell output), which are multiplied with weight matrices before bias is added. The output of the activation function, which receives the outcome, is binary. When a cell state's output is 0, the piece of information is lost, but when it is 1, the information is saved for use later.

2) Input Gate

The addition of useful information to the cell state is done by the input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered like the forget gate using inputs h_{t-1} and x_t. Then, a vector is created using the tanh function that gives an output from -1 to +1, which contains all the possible values from h_{t-1} and x_t. At last, the values of the vector and the regulated values are multiplied to obtain useful information.

3) Output gate

The output gate's job is to take meaningful information out of the current cell state and deliver it as output. The tanh function is first used on the cell to create a vector. The data is then filtered by the values to be remembered using the inputs h_{t-1} and x_t, and the information is then controlled using the sigmoid function. The vector's values and the controlled values are finally multiplied and supplied as input and output to the following cell, respectively.

V. RESULTS AND DISCUSSION

In this system, an end-to-end web application from scratch has been developed. Basically, there are three modules in the system. In Module I, Data Generation and Collection, Web application Design and Development, VisObjects Development and Role based Security has been implemented. In Module II, Topic Modeling has been implemented using ML and NLP Techniques which includes S-BERT, UMAP and HDBScan. In Module III Sales Prediction has been implemented using Supervised Machine Learning Algorithms including Linear Regression, ARIMA, Random Forest, XGBoost, LSTM.

A. Module I

*1) Login Page*An web page or user interface known as a login page enables users to log in to a secure system, application, or website by providing their login information, such as a username and password. It functions as an authentication method to confirm users' identities and grant them the proper access privileges.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

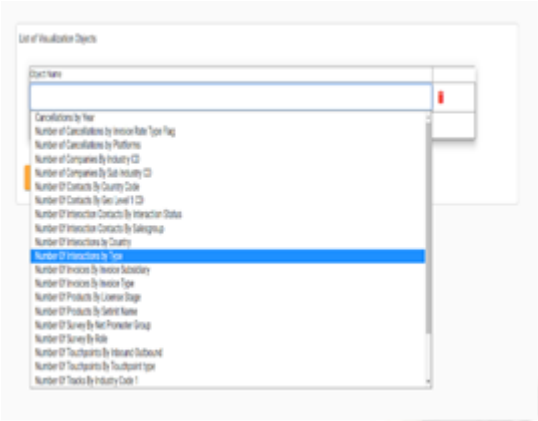

Figure 16. Login Page (Web Application).

2) List of VisObjects

A list of VisObjects is a group of visualization tools used for data visualization or graphical data representation. These items are frequently pieces of software that facilitate the presentation of data visualizations. They offer a more natural and approachable way to analyze and convey complicated data correlations, trends, and insights using visual means. With the aid of VisObjects, users may analyze data visually, come to wise judgements, and effectively communicate.

Figure 17. List of VisObjects.

3) List of Reports

A list of reports or dashboards is a collection of different VisObjects, such as graphs, charts, and other visual elements, that work together to offer an extensive and interactive overview of data and information.

	text	label_st1
0	realli like game repetit user friend	38
1	hate control	17
2	pretti need lot work vr titl moment system dunki enough moter control option feel fight game rather play hope patch buy anoth game vr support skip get patch # # vr system use vive # #	59
3	/ - better take detour around game neat idea neat present neat graphicsand thing neat " - get bore pretti quick	69
4	ok lot repetit realli recommend go play dungeon keeper real experi even graphic origion dungeon keeper beat game play enjoy	-1

Figure 18. List of Reports.

Users can conduct comprehensive and insightful data analysis using these reports or dashboards, which act as visual representations of important metrics, trends, and insights.

B. Module II

1) Clustering DF

A Dataframe for clustered textual review data organizes and represents textual reviews according to the clusters or groups to which they have been assigned. Dataframe's properties include the actual review content, cluster labels designating the group to which the review belongs, and additional data such as review scores.

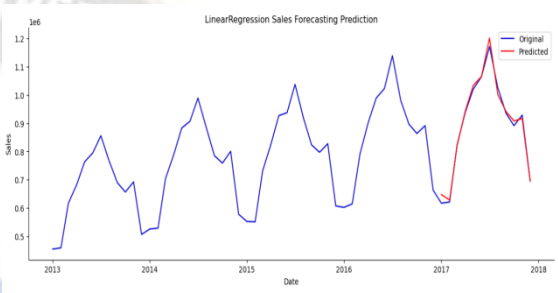

Figure 19. Dataframe for Clustered Data.

C. Module III

In Module III of the sales forecasting module, Supervised algorithms has been applied to predict sales. Linear Regression, Random Forest, XGBoost, LSTM, and ARIMA are the various algorithms which has been used.

1) Sales Prediction Using Various Supervised ML Techniques

a) Linear Regression

Linear Regression is a straightforward algorithm that assumes a linear relationship between the input variables and the target variable. It estimates coefficients to minimize the difference between predicted and actual values, making it suitable for cases with a linear relationship.

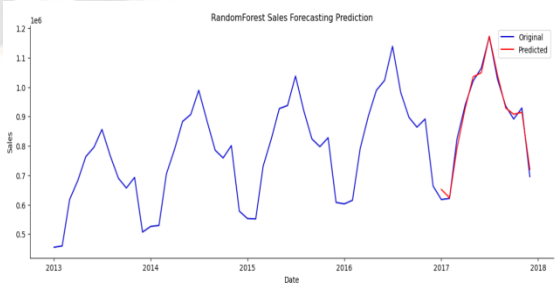

Figure 20. Sales Prediction using Linear Regression.

b) Random Forest

Random Forest Regression, an ensemble learning algorithm, combines multiple decision trees to make

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

predictions. It handles non-linear relationships and interactions among variables effectively, making it a powerful tool for sales prediction.

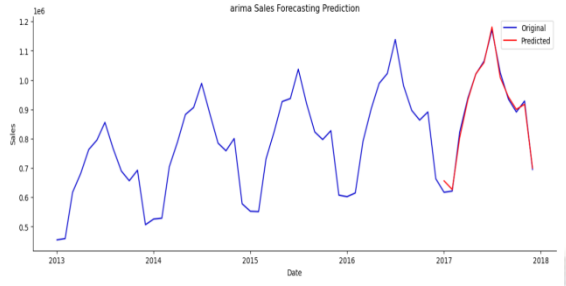

Figure 21. Sales Prediction using Random Forest.

c) ARIMA

ARIMA (Auto Regressive Integrated Moving Average) is a statistical method that models the temporal structure of time series data. It considers auto-regression (dependence on previous observations), differencing (to achieve stationarity), and moving averages (smoothing out noise) to forecast sales. A Machine Learning technique for time series forecasting that uses historical sales data to estimate future sales is called ARIMA. It simulates the variance in the forecast and the link between an actual sales series and its own lag values. Autoregression (AR), differencing (I), and moving average (MA) are the three main factors that ARIMA considers. It analyses the patterns and trends in the data to forecast future sales values, offering businesses a useful tool to foresee sales variations, schedule inventory levels, and reach well-informed decisions.

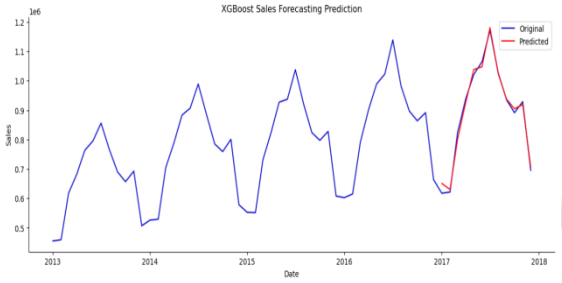

Figure 22. Sales Prediction using ARIMA.

d) XGBoost

XGBoost, an optimized gradient boosting algorithm, utilizes an ensemble of weak prediction models. It excels at capturing complex patterns and non-linear relationships in the data, making it suitable for accurate sales forecasting.

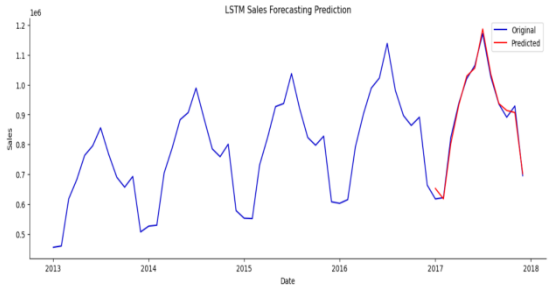

Figure 23. Sales Prediction using XGBoost.

e) LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is capable of learning long-term dependencies. LSTMs are particularly useful when dealing with sequential sales data, capturing temporal patterns and dependencies.

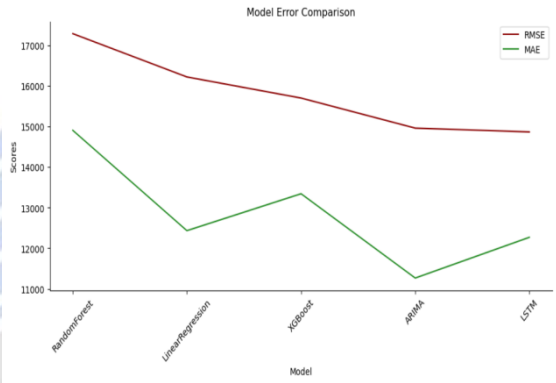

Figure 24. Sales Prediction using LSTM.

2) RMSE and MAE Graph

Based on the RMSE and MAE values, the lower the values, the better the model's performance in terms of accuracy. Comparing the models, the Random Forest model has the highest RMSE value and the second highest MAE value, indicating that it has higher prediction errors compared to the other models.

On the other hand, the LSTM model has the lowest RMSE value and the second lowest MAE value, suggesting that it performs better in terms of accuracy compared to the other models. The ARIMA model also performs well with relatively low RMSE and MAE values.

Based on these results, it can be inferred that the LSTM and ARIMA models are likely to provide more accurate sales forecasts compared to the Random Forest, Linear Regression, and XGBoost models. However, it is important to consider other factors such as computational complexity, interpretability, and the specific requirements of the forecasting task before selecting the most suitable model for implementation.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

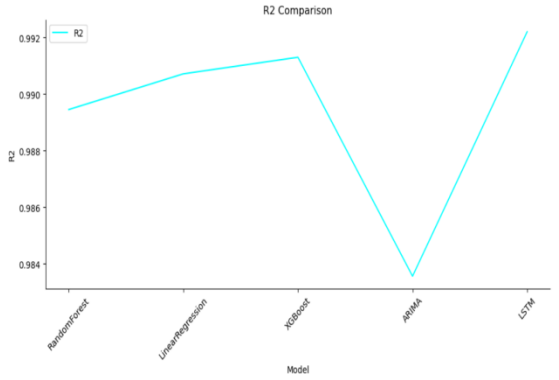

Figure 25. RMSE amd MAE Comparison.

3) R2 Graph

R2 (R-squared) values indicate the goodness-of-fit of a regression model. The closer the R2 value is to 1, the better the model fits the data. Comparing the provided R2 values, the highest R2 value is observed for the LSTM model, indicating that it has the best fit to the data among the given models.

Following closely behind are the XGBoost and Linear Regression models respectively. These models also demonstrate a strong fit to the data. The Random Forest model has a slightly lower R2 value, but it still shows a good fit to the data.

The ARIMA model has the lowest R2 value, suggesting a relatively weaker fit compared to the other models. However, it is important to note that ARIMA is a time series forecasting model and its performance is often evaluated based on other metrics such as RMSE and MAE.

Based on the R2 values, the LSTM model seems to be the most suitable for the given sales forecasting task, closely followed by XGBoost and Linear Regression. However, it is important to consider other factors such as interpretability, computational complexity, and specific requirements of the task when choosing the most appropriate model for implementation.

VI. CONCLUSION

In conclusion, the implementation of a business analytics system using predictive algorithms and natural language processing techniques offers valuable solutions for sales forecasting and extracting insights from unstructured textual data.

Firstly, in Module I, web application development has been implemented using Spring Boot, Angular, and MySQL. The integration of these technologies allowed for the creation of a functional web application with a responsive front-end and a robust back-end database management system. Secondly, in Module I, the reports have been designed using various visualization objects using the Chart JS library. This enabled the representation of data in a visually appealing and informative manner, facilitating data analysis and decision-making processes.

Moving on to Module II, the topic modeling task involved clustering using the unsupervised algorithm HDBScan. The algorithm identified clusters within the dataset, providing insights into patterns and relationships among the data points. The resulting graph of clusters helped visualize the identified groups and their interconnections.

In Module III, sales forecasting has been conducted using various supervised algorithms such as Linear Regression, Random Forest Regression, XGBoost, LSTMs, and ARIMA. The comparison of MAE and RMSE metrics provided an evaluation of the algorithms' performance. Lower MAE and RMSE values indicated more accurate sales predictions. Experimentation has been carried out to determine the performance of the techniques, which yields R2 values 0.9894, 0.9907, 0.9913, 0.9835, and 0.9922 of RFR, LSTM, LR, XGBoost and ARIMA respectively.

Overall, the system demonstrated the successful implementation of web application development, visualization, topic modeling, and sales forecasting. The integration of these modules showcased the potential for data-driven decision-making, enabling businesses to make informed choices based on accurate predictions and insights derived from the data.

REFERENCES

[1] G. D. Kalyankar, S. R. Poojara, and N. V. Dharwadkar, "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Feb. 2017, doi:10.1109/I-SMAC.2017.8058253.

[2] I. M. Al-Zuabi , A. Jafar and K. Aljoumaa, "Predicting customer's gender and age depending on mobile phone data," Journal of Big Data, 2019, doi:10.1186/s40537-019-0180-9.

[3] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," Journal of Big Data, 2019, doi:10.1186/s40537-019-0175-6.

[4] A. R. Kaufman, P. Kraft and M. Sen, "Improving Supreme Court Forecasting Using Boosted Decision Trees," Political Analysis, vol. 27, issue 3, Cambridge University Press, Jul. 2019, pp. 381-387, doi:10.1017/pan.2018.59.

[5] I. Purnamasari, F. Handayanna, E. Arisawati, L. S. Dewi, E. G. Sihombing and Rinawati. "The Determination Analysis Of Telecommunications Customers Potential Cross- Selling

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7242
***Article Received**: 28 April 2023 **Revised:** 24 June 2023 **Accepted:** 02 July 2023*

With Classification Naive Bayes And C4.5," International Conference on Advanced Information Scientific Development (ICAISD), vol. 1641, Aug. 2020, doi:10.1088/1742-6596/1641/1/012010.

[6] N. B. a-Misu and M. Madaleno, "Assessment of Bankruptcy Risk of Large Companies: European Countries Evolution Analysis," Journal of Risk and Financial Management, vol. 13, issue 3, 2020, doi:10.3390/jrfm13030058.

[7] A. Tolba and Z. Al-Makhadmeha, "Predictive data analysis approach for securing medical data in smart grid healthcare systems," Future Generation Computer Systems, vol. 117, Apr. 2021, pp. 87-96, doi:10.1016/j.future.2020.11.008.

[8] P. Punjabi, P. Vaswani and A. Kubal, "Modelling Stock Trading Platforms Leveraging Predictive Analysis Using Learning Algorithms," IJRAR - International Journal of Research and Analytical Reviews (IJRAR), vol. 8, issue 2, pp. 422-438, Jun. 2021, ssrn:https://ssrn.com/abstract=3868081.

[9] A. Saleh Hussein, R. Salah Khairy, S. M. Mohamed Najeeb, and H. T. Alrikabi, "Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression," International Journal of Interactive Mobile Technologies (iJIM), vol. 15, no. 05, pp. 24–42, Mar. 2021, doi: 10.3991/ijim.v15i05.17173.

[10] S. Sawangarreerak and P. Thanathamathee, "FakeBERT: Detecting and Analyzing Fraudulent Patterns of Financial Statement for Open Innovation Using Discretization and Association Rule Mining," Journal of Open Innovation: Technology, Market, and Complexity, vol. 7, issue 2, 2021, doi:10.3390/joitmc7020128.

[11] M. Seera, C. P. Lim, A. Kumar, L. Dhamotharan and K. H. Tan, "An intelligent payment card fraud detection system," Annals of Operations Research, Jun. 2021, doi:10.1007/s10479-021-04149-2.

[12] S. Ayvaz and K. Alpay, "Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time," Expert Systems with Applications, vol. 173, Jul. 2021, doi:10.1016/j.eswa.2021.114598.

[13] D. Huo, H. R. Chaudhry, "Using machine learning for evaluating global expansion location decisions: An analysis of Chinese manufacturing sector," Technological Forecasting Social Change, vol. 163, Feb. 2021, doi:10.1016/j.techfore.2020.120436.

[14] M. Kumar, V. M. Shenbagaraman, R. Shaw, and A. Ghosh, "Predictive Data Analysis for Energy Management of a Smart Factory Leading to Sustainability," Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering, vol 661, Jan. 2021, pp. 765-773, doi:10.1007/978-981-15-4692-1_58.

[15] E. Ruschel, E. De F. R. Loures and E. A. P. Santos, "Performance analysis and time prediction in manufacturing systems," Computers & Industrial Engineering, vol. 151, Jan. 2021, doi:10.1016/j.cie.2020.106972.

[16] C. S. Lee, P. Y. S. Cheang, M. Moslehpour, "Predictive Analytics in Business Analytics: Decision Tree," Advances in Decision Sciences, vol. 26, issue 1, 2022, pp. 1-30, doi:10.47654/v26y2022i1p1-30.

[17] M. Al-Omari, F. Qutaishat, M. Rawashdeh, S. H. Alajmani and M. Masud, "A Boosted Tree-Based Predictive Model for Business Analytics," Intelligent Automation Soft Computing, vol.36 , issue 1, 2022, doi:10.32604/iasc.2023.030374.

[18] M.A. Waller and S. E. Fawcett, "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management", Journal of Business Logistics, vol. 34, no. 2, Jun 2013, pp. 77-84, doi:10.1111/jbl.12082.

[19] G. Shmueli, and O.R. Koppius, "Predictive analytics in information systems research," MIS Quarterly, vol.35, no. 3, Sep. 2011, pp. 553-572, doi:10.2307/23042796.

[20] D. E. Brown, A. Abbasi, and R. Y. K. Lau, "Predictive analytics: Predictive modeling at the micro level", IEEE Intelligent Systems, vol. 30, no. 3, May 2015, pp. 6-8, doi:10.1109/MIS.2015.50.

[21] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, vol. 35, no. 2, Apr. 2015, pp. 137-44, doi:10.1016/j.ijinfomgt.2014.10.007.

[22] D. Larson, and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science", International Journal of Information Management, vol. 36, no. 5, Oct. 2016, pp.700-710, doi:10.1016/j.ijinfomgt.2016.04.013.